Modelo determinista para el análisis sintáctico: una perspectiva algorítmica y computacional de la lingüística

Julio Meroño Sáez

8 de enero de 2024

Resumen

(English version)

(Versión en español)

Índice

Parte I

Prólogo

1. Introducción

El español o castellano es una lengua iberorromance y es la lengua nativa de más de 450 millones de hablantes en el mundo, a lo que se le suman los 75 millones de hablantes no nativos alrededor del planeta. Esto la sitúa como la cuarta lengua más hablada globalmente, detrás del inglés, el chino mandarín y el hindi. Es por ello que la existencia de una lingüística

estandarizada es fundamental, al permitir la comunicación internacional sin cabida a ningún tipo de ambigüedad y al mantener la integridad de la lengua. La Asociación de Academias de la Lengua Española, junto a sus correspondientes academias nacionales (entre ellas, la Real Academia Española), son las encargadas de dicho cometido, proporcionando una estandarización de la gramática, morfología, léxico y fonética del español alrededor del mundo hispanohablante.

Sin duda, la gramática es tan fundamental como las otras ramas de estudio de la lingüística. Esta es la encargada de estudiar la organización y la función de las palabras dentro de una oración, junto a la agrupación de palabras en unidades mayores con significado y función, denominadas sintagmas. A su vez, dichos sintagmas se organizan por categorías sintagmáticas que engloban y marcan su función general. El estudio y el estandarizado de la gramática en el español se remonta a finales del siglo XV, con autores como Antonio de Nebrija, llegando y continuando hasta la actualidad con teorías y propuestas transversales a la mayoría de lenguas, como las propuestas por Noam Chomsky, o trabajos más enfocados en el español.

Tradicionalmente, y aún en épocas recientes, se ha realizado un estudio sintáctico más específico para cada categoría sintagmática, donde los sintagmas (o grupos, como se hacen llamar en ciertos modelos) poseen estructuras fundamentalmente distintas entre tipos, y donde se da lugar a numerosos casos límite a tener en cuenta. Aún así, dicho análisis ha resultado ser considerablemente más efectivo y simple a la hora de su uso superficial por lingüistas procedentes de otras ramas, o en el ámbito educativo como medio de enseñanza. Sin embargo, desde finales del siglo anterior y gracias a la teoría estándar propuesta por Noam Chomsky, se ha dado lugar a otro enfoque distinto a la gramática, denominado gramática generativa. En la mayoría de teorías y modelos asociados con dicha gramática, se formula que todas las agrupaciones a las que se les denomina sintagmas comparten una estructura fundamental, que aun pudiendo variar en el tipo de unidades y tipos de sintagmas agrupados para formarlo, siempre se estructuran de la misma manera, sin casi ninguna cabida a variación.

El estudio de la gramática en el español, especialmente de la sintaxis, ha resultado en el creciente interés por el uso de las nuevas tecnologías y de la capacidad computacional de este siglo para el análisis sintáctico de frases, sintagmas individuales y oraciones. Por

desgracia, el uso en su mayoría de la gramática tradicional para la creación de modelos y algoritmos que cumplan dicha tarea dificulta considerablemente la detección y el correcto análisis de ciertos casos especiales o límite que se pueden dar, o imposibilita casi por completo el proporcionamiento de más información en cuanto a la función que realizan ciertos sintagmas dentro de agrupaciones mayores, factor que la gramática tradicional no suele tener en cuenta.

Es por ello por lo que este trabajo de investigación propone un análisis determinista en forma de modelo adaptado de la gramática del español, basándose fundamentalmente en la gramática generativista aportada en la obra de Karen Zagona, Sintaxis generativa del español [1], pero incluyendo modificaciones tanto propias como de otros autores que permitan una mayor facilidad para el procesado de oraciones mediante implementaciones computacionales de dicho modelo. Además, se pretende diseñar y llevar a cabo una implementación tipo del modelo formulado previamente que sea capaz de poder utilizarse en ordenadores convencionales sin requerir de un excesivo coste computacional, para su posible y posterior publicación mediante una interfaz en línea y de acceso libre para su uso público.

2. Justificación

Son decenas de millones los estudiantes que se encuentran actualmente aprendiendo español en las escuelas. En específico, contando únicamente los estudiantes extranjeros y no procedentes de ningún país hispanohablante, se estima que el número se encuentra en el orden de los 24 millones [2]. Entre toda la materia y contenidos aportados en la clase a dichos estudiantes durante sus horas lectivas, una gran y significativa parte consiste en el estudio de la propia lengua. Además, la utilización del estudio y análisis sintáctico mediante separación en sintagmas o grupos encuadrados se ha vuelto a lo largo de los años el estándar a nivel nacional y, en ciertos casos, internacional. Aunque los estudios que lo afirmen son relativamente escasos, predominantemente la experiencia ha demostrado a las instituciones que planifican las competencias a tener en cuenta a la hora de enseñar en las aulas que el entendimiento del funcionamiento de la lengua implica un mejor uso de esta, tanto social y oral como formal y escrito, y es fundamental para la preservación del español, razón junto a las anteriores por las que se muestra un interés significativo en su enseñanza en los centros

educativos a lo largo de todo el territorio español. Además, se ha de tener en cuenta la reciente y creciente adopción de esta metodología en otros países de habla española, lo que implica una mayor impulsión del conocimiento interno del funcionamiento de la sintaxis española.

Si bien, el uso de la gramática tradicional como herramienta didáctica se remonta a, aproximadamente, el siglo anterior, la cantidad de material de ejercicios para los estudiantes es particularmente escaso, especialmente cuando se pretende comparar con otras materias donde no solo existe mucho más material, sino que es posible su generación procedural y automática. En principio, lo comentado no debería tener ninguna implicación negativa, pues al tratarse de un estudio de la gramática española en forma de sintaxis basada en teorías tradicionales, la generación de ejercicios resulta tan simple como obtener, buscar o producir por la propia cuenta del profesorado oraciones que se adapten al nivel de los estudiantes. Sin embargo, se ha de considerar y tener en cuenta que, para comprobar la correctitud y validez de las respuestas otorgadas por los estudiantes a espera de ser evaluadas, resulta tedioso tener que realizar una comprobación mental lenta y laboriosa de dichos resultados, que puede además dar lugar a numerosos errores de comprobación, no solo beneficiosos sino también perjudiciales en ciertos casos para el alumnado. Por otro lado, desde el punto de vista estudiantil, la comprobación independiente de resultados es completamente imposible, pues se carece de un punto de referencia objetivo para comparar, por lo que estos carecen de una forma de poder practicar de manera efectiva e ilimitada, sin depender del continuo trabajo del profesorado.

Lo que diferencia principalmente a las personas de cualquiera de los dos puntos de vista anteriores con los ordenadores y dispositivos electrónicos es la capacidad de poder tomar decisiones aplicando el entendimiento horizontal, subjetivo y basado en la experiencia y no en reglas a aplicar. Por lo tanto, la tarea de producir análisis sintácticos utilizables en el ámbito de la creación de asistentes basados en la inteligencia artificial se considera un tema popular y continuamente en desarrollo y debate en investigaciones de las últimas décadas.

Este último caso nos lleva al ámbito judicial, donde se muestra una carencia de estandarización intuitiva del significado de los textos legislativos, los contratos y los acuerdos formales, dando lugar a que existan profesiones cuyo único propósito sea el de garantizar la ausencia de vacíos

legales a la hora de producir dichos papeles. Todo esto ocurre por una vaga especificación abierta a interpretaciones de la gramática utilizada en este campo.

La motivación de este trabajo de investigación es, entonces, aportar un pequeño pero significante grano de arena a la base de lo que será el mundo en los años y las décadas por venir, tanto en el ámbito de la educación, como en el de la tecnología, como en el judicial, como en cualquier otro que se pueda plantear (lenguaje matemático, diplomacia, etcétera). Para ello, se plantean diversos objetivos escalonados que se formulan en el apartado a continuación.

3. Objetivos del trabajo

El objetivo principal de este trabajo de investigación es, en resumidas cuentas, poder aportar un modelo descriptivo y determinista pero empírico de la gramática del español, basándose en una agrupación de teorías y definiciones de la gramática de autores anteriores y antecedentes, mezclado con un enfoque propio matizado por la mentalidad computacional y matemática, que pueda utilizarse como herramienta de formalización, educación y comprobación, pero además como punto de partida para futuras investigaciones.

Por otro lado, este objetivo ya mencionado da lugar a ciertos enfoques alternativos pero compatibles por los que llevar la investigación, en forma de objetivos secundarios:

1. Crear una herramienta de uso fácil e inmediato para el análisis de oraciones y sintagmas en español, disponible de forma completamente gratuita al público, para su uso por profesorado, estudiantes, políticos, diplomáticos, abogados, etcétera.

2. Aportar una simplificación general de las estructuras de árbol presentes en la mayoría de corrientes presentes, principalmente del árbol bipartito propuesto por las principales ramas de la gramática generativista, que permita su uso de forma más eficaz a la hora de realizar análisis computacionales de la sintaxis del español, no solo como punto de apoyo para el anterior objetivo sino también como base para modelos de asistencia mediante inteligencia artificial, chatbots y cualquier tecnología similar.

3. Agrupar los esfuerzos provenientes de las distintas corrientes lingüísticas existentes para

propulsar la gramática española en un trabajo recopilatorio que, además, introduce ciertas variaciones partiendo de un enfoque más tecnológico, algorítmico y modélico.

Parte II

Modelo adaptado de la gramática española

En esta parte del trabajo de investigación, se diseña y redacta un modelo adaptado de la gramática del español, con la intención de su posterior uso en la implementación del mismo mediante técnicas propias del ámbito de la programación y la algoritmia. Previamente a ello, sin embargo, se nombra la base a partir de la que se construye el modelo, y se menciona en detalle la estructura, metodología de diseño y paradigma utilizado para la creación del mismo.

Es de relevancia para el entendimiento del trabajo tener en cuenta que, en la redacción del modelo, se parte de una base nula, posiblemente reasignando definiciones incompatibles a cierta terminología en uso con el fin de simplificar la producción y aumentar la claridad de este. Además, las propuestas de clasificación, análisis y estructurado pueden poseer disparidades significativas con las procedentes de las corrientes gramaticales actuales, incluso, en definiciones generalmente aceptadas previamente. Esto es debido, por otro lado, a los intentos de acercar el modelo a una redacción lo suficientemente formal en el sentido algorítmico como para su utilización sin posible lugar a dudas en las implementaciones.

Por último, por limitaciones de tiempo, de personal y de antecedentes previos a este trabajo de investigación, resulta considerablemente dificil, incluso imposible en la práctica, tratar de definir la totalidad de la gramática española en el mismo, por lo que se ha decidido reducir la lengua en cuestión a un subconjunto del español. Con esto se pretende decir, adicionalmente, que cualquier enunciado válido en esta versión reducida del español es, también, un enunciado válido en el español real. Esta reducción elimina ciertas estructuras complejas, como la gran mayoría de construcciones oracionales o la yuxtaposición de cualquier tipo.

También se descartan ciertas variaciones internas dentro las estructuras sintagmáticas, limitándolas a formas y ordenamientos fácilmente procesables y comunes en el español, mencionados posteriormente en el trabajo.

4. Fuentes e investigaciones previas

El modelo presentado y definido a lo largo de las siguientes secciones obtiene su inspiración y punto de partida en diversas corrientes gramaticales y lingüísticas existentes en el momento. Sin embargo, dichas corrientes carecen del rigor matemático o algorítmico necesario para la realización de implementaciones formalmente verificables. De esta manera, el marco bibliográfico constituido por trabajos de relevancia para el diseño del modelo se encuentra severamente limitado. Esto conlleva, como se ha mencionado previamente, la necesidad de simplificar y reducir la gramática tratada en el modelo a un conjunto reducido y delimitado por lo fácilmente deducible y extrapolable a definiciones formales de lo explicado en las fuentes bibliográficas e investigaciones previas de este trabajo de investigación. Sin embargo, esto implica que las fuentes en uso juegan un papel meramente inspiracional o de guía en las definiciones aportadas, por lo que se ruega discreción a la hora de su interpretación en las secciones posteriores.

A continuación, se mencionan brevemente las fuentes que se han utilizado para la construcción y el diseño del modelo:

1. *Corpus AnCora.* [1] Este corpus español-catalán posee diversas anotaciones morfológicas, sintácticas y semánticas de significativa utilidad para este trabajo de investigación. En especial, los dos lexicones extraídos del mismo, AnCora-Verb y AnCora-Nom, forman la base del procesado semántico efectuado en la implementación del modelo. Las agrupaciones en argumentos de los roles semánticos presentes en el modelo se realizan con la intención de simplificar la tarea de llevar su análisis a la práctica, alineándolas con las presentes en estos lexicones.

2. *Nueva gramática de la lengua española.* [2] Desde su publicación en 2009, esta obra académica ha servido como actualización de la gramática normalizada por las mismas entidades en la primera mitad del siglo anterior. En este trabajo de investigación,

la gramática procedente de la misma ha constituido la base para gran parte de la terminología y definiciones en uso y la clasificación de elementos, especialmente en los apartados más relacionados con la morfología y la enumeración de rasgos gramaticales.

3. *Corpus del español del siglo XXI.* [3] Las tablas de frecuencias de este corpus han sido de considerable utilidad para la formación de la base de datos de elementos gramaticales en uso en la posterior implementación, además de proporcionar un análisis de rasgos gramaticales significativamente exhaustivo, razón por la que esta fuente ha sido crucial para el trabajo de investigación. Es por ello por lo que, para el diseño del modelo, se ha tomado la leyenda de rasgos gramaticales proporcionados en el análisis como punto de partida para su nombramiento y clasificación.

4. *Sintaxis generativa del español.* [4] En este libro, se proporciona una visión global de la sintaxis del español desde un punto de vista basado en la gramática generativista. De esta visión surge gran parte de las estructuras de las cláusulas y los sintagmas tratadas en este trabajo de investigación, además de, en general, servir como inspiración movida por las corrientes generativistas para la planificación y el diseño de ciertas secciones del modelo, tanto en los apartados enfocados en la morfología como en los que tratan principalmente la sintaxis del subconjunto del español en cuestión.

5. Estructura y visión del modelo

Dentro del modelo redactado a continuación, se proporciona un enfoque mayoritariamente funcional y empírico, evitando comprometer la simplicidad del modelo que se perdería al pretender tratar temas como el origen natural de ciertas estructuras. Por ello, se proporcionan declaraciones que repliquen las construcciones naturales de la lengua con la mayor precisión posible dentro de los límites prácticos del trabajo.

La estructura del mismo consiste en tres niveles fundamentales, ordenados según su dependencia en los niveles anteriores y según el grado de abstracción presente en cada uno de ellos. De esta manera, se comienza tratando el nivel léxico-morfológico, donde se trata la base léxica de la que se parte, sus rasgos gramaticales y la clasificación de estos. A continuación,

se menciona el nivel sintáctico, en el que se explican las distintas estructuras válidas para la agrupación de lexías en enunciados o subdivisiones de los mismos. Por último, se describe el nivel semántico, cuya función es la de extraer y procesar el significado de las estructuras formadas mediante la asignación de roles.

De cualquier manera que se plantease el trabajo, sin embargo, siempre sería necesario partir de una base de conocimientos previos no tratados en el mismo, por lo que, durante la creación del modelo, ciertos apartados han sido omitidos por cuestiones de limitaciones o de conveniencia. Específicamente, dos casos son inmediatamente perceptibles y de relevancia:

1. El nivel léxico-morfológico presenta un vacío de contenido en cuanto al origen, la formación y la flexión de elementos gramaticales, pues, al partir del análisis exhaustivo de elementos previamente realizado por la Real Academia Española, resulta totalmente innecesario definir el proceso a llevar a cabo para efectuar dicho análisis. Así, se ignora en el trabajo de investigación la relación existente entre la organización de los morfemas en los elementos y los rasgos gramaticales que surgen como producto de esta organización.

2. El nivel semántico utiliza una agrupación de roles en argumentos que, aun lógica y válida a la vista, resulta aparentemente arbitraria en cuanto a su selección. Dicha agrupación proviene, sin embargo, de un intento de compatibilizar los argumentos definidos con los presentes en los lexicones procedentes del corpus *AnCora*, posteriormente utilizados en la implementación de este nivel del modelo.

Por último, a lo largo de toda la redacción que abarca el modelo a continuación, se utiliza en las secciones una notación específica para definir ciertos apartados o estructuras formalmente. Para evitar cualquier tipo de confusión, y definirla en su totalidad en los casos donde esta difiera de la norma, se mostrarán ejemplos y explicaciones previas a su utilización en el trabajo.

6. Nivel léxico-morfológico

El nivel léxico-morfológico es el primer nivel a tratar en el modelo, pues trata la base léxica sobre la que se contruyen, posteriormente, las estructuras internas de los enunciados.

Además, se describen todos los posibles rasgos de las unidades gramaticales, con el fin de poder utilizarlos para hacer posible su análisis en el nivel sintáctico y en el nivel semántico. Para ello, se introduce la lexía como unidad elemental sobre la que construir el resto del modelo.

La lexía es un elemento gramatical cuyas partes presentan un elevado índice de inseparabilidad morfológica o forman una función distinta, no necesariamente presentando cierto valor semántico, que cualquier subconjunto de dichas partes. Las lexías engloban términos más convencionales como la palabra y les aporta una definición menos problemática, pero no se limitan únicamente a estos.

Según su estructura morfológica y visual, las lexías se pueden clasificar de la siguiente manera:

1. *Lexías simples.* Constan de un conjunto de morfemas carente de espaciado o separación simbólica alguna. Se incluyen aquí tanto formas simples como compuestas de las palabras, a diferencia de la clasificación convencional de lexía, pues permite un procesado más simple de las mismas.

2. *Lexías complejas.* También denominadas como locuciones, se componen de varios morfemas donde sí se presenta una separación mediante espacios o símbolos. Cada isla de morfemas se podría considerar equivalente a una lexía simple, sin embargo, la totalidad de la lexía aporta una función y un valor semántico distintos a cada una de las lexías simples discernibles en su interior, o cualquier agrupación de estas.

3. *Lexías textuales.* Los nombres propios, los términos numéricos, las expresiones hechas, las citaciones y cualquier término insondable por las herramientas del ámbito gramatical se consideran lexías textuales, pues difieren de las lexías simples y las lexías complejas en que su significado proviene de una asignación artifical del mismo, de un contexto externo al entorno o de una interpretación poética de una o más lexías de otra clase.

4. *Símbolos.* Estos son constituidos por cualquier marcación indivisible pero estrictamente individual con una función exclusivamente gramatical. Sus usos incluyen, pero no se

limitan a, la delimitación o la enumeración de elementos gramaticales, y no constan en ningún alfabeto.

Por otro lado, según su función y la presencia de valor semántico, se pueden clasificar como se presenta a continuación, siendo esta la clasificación de mayor utilidad a lo largo del modelo:

1. *Lexías semánticas.* Se diferencian al poseer un valor semántico y al constituir casi la totalidad del repertorio léxico. Los adjetivos, los adverbios, los sustantivos y los verbos son los cuatro tipos de lexías semánticas.

2. *Lexías determinantes.* Tienen como función la de concretar o determinar lexías, tornando conceptos en referencias. Además, pueden aportar rasgos gramaticales adicionales a dichas lexías, o reforzar los rasgos ya presentes en ellas.

3. *Lexías gramaticales.* Estos cohesionan diversas lexías entre sí, formando jerarquías y enumeraciones, o sirven para cambiar la función de otras lexías. Se incluyen aquí las conjunciones, las preposiciones y los símbolos.

6.1. Rasgos gramaticales

Los rasgos gramaticales son propiedades inherentes de las lexías que pueden tomar valores cuantizados y limitados.

6.2. Las lexías semánticas

Adjetivos

Adverbios

Sustantivos

Verbos

6.3. Las lexías determinantes

Artículos

Cuantificadores

Demostrativos

Interrogativos

Numerales

Posesivos

Pronombres personales

Relativos

6.4. Las lexías gramaticales

Conjunciones

Preposiciones

Símbolos

7. Nivel sintáctico

7.1. La enumeración

7.2. La cláusula

7.3. El sintagma

El sintagma adjetival

El sintagma adverbial

El sintagma nominal

El sintagma verbal

8. Nivel semántico

Parte III

Implementación del modelo

aaa pero AAA, asi que aaa

9. Accesibilidad al público

10. Evaluación del modelo

11. Conclusiones

Referencias

Antònia Martí, M., Taulé, M., Bertran, M., & Màrquez, L. (2007). AnCora: Multilingual and Multilevel Annotated Corpora. https://clic.ub.edu/corpus/es/ancora

Asociación de Academias de la Lengua Española [ASALE], Real Academia Española [RAE]. (2009). *Nueva gramática de la lengua española*. Espasa-Calpe.

Real Academia Española [RAE]. (2023). Corpus del Español del Siglo XXI. https://www.rae.es/corpes

Zagona, K. (2006). *Sintaxis generativa del español*.